

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-108887

(43)Date of publication of application : 12.04.2002

(51)Int.Cl. G06F 17/30

(21)Application number : 2000-299971 (71)Applicant : CANON INC

(22)Date of filing : 29.09.2000 (72)Inventor : NAKAKOSHI HARUKI

**(54) DOCUMENT RETRIEVERMETHOD FOR ADDING KEYWORD TO THE
RETRIEVERDOCUMENT RETRIEVAL METHOD AND COMPUTER READABLE
STORAGE MEDIUM**

(57)Abstract:

PROBLEM TO BE SOLVED: To attain the effective retrieval of a document even to an unknown keyword.

SOLUTION: When a character string including a set keyword is retrieved out of a document to be retrieved in accordance with the keyworda relation dictionary is referred to according to the set keyword for setting additionally the word that is related with the keyword as another keyword and a character string including at least one of both keywords is retrieved out of the document to be retrieved in accordance with these keywords (S11-S15). When the word extracted from an input document is equal to an unknown word that is not registered on a previously stored keyword dictionarythe relation dictionary relates the extracted word with the word that is positioned before or after the unknown word of the input document and is produced before the steps S11-S15.

CLAIMS

[Claim(s)]

[Claim 1]A document retrieval system with which a character string containing the keyword is searched out of a document of a retrieval object based on a set-up keywordcomprising:

An unknown word detection means to judge whether it is an unknown word by which a word which extracted a word and was extracted is not registered into the 1st dictionary from an input sentence document.

A word judged to be an unknown word by said unknown word detection means.

A correlation means to draw up the 2nd dictionary in which a word located the front or behind this word was associated. [in said input sentence document]

A keyword adding means which sets additionally a word related with this keyword by referring to said 2nd dictionary based on a set-up keyword as a new keyword.

[Claim 2]The document retrieval system according to claim 1 having a search means to search a character string containing at least one of these keywords out of a document of said retrieval objectaccording to two or more keywords developed by said keyword adding means.

[Claim 3]The document retrieval system according to claim 2 when said search means is not detected [an unknown word] by said unknown word detection meanswherein it searches only a character string containing said set-up keyword out of a document of said retrieval object.

[Claim 4]The document retrieval system according to claim 1 or 2 which is characterized by a commuter's ticket or performing irregularly to a document of said retrieval object separately from execution timing of said keyword adding means and said search means as for said unknown word detection means and said correlation means.

[Claim 5]The document retrieval system according to claim 1wherein said keyword adding means contains a means to set these keywords additionally furtherbased on said set-up keyworda keyword added by referring to said 2nd dictionaryand said 1st dictionary.

[Claim 6]They are the keyword additional methods to a document retrieval system with which a character string containing the keyword is searched out of a document of a retrieval object based on a set-up keywordAn unknown word detection process which extracts a word and judges whether it is an unknown word by which an extracted word is not registered into the 1st dictionary from an input sentence documentA correlation process of drawing up the 2nd dictionary in which a word judged to be an unknown word in said unknown word detection process and a word located the front or behind this word were associated[in said input sentence document] Keyword additional methods having like keyword additional processing which sets additionally a word related with this keyword by referring to said 2nd dictionary based on a set-up keyword as a new keyword.

[Claim 7]A document retrieval method searching a character string containing at least one of these keywords out of a document of said retrieval object according to two or more keywords developed with the keyword additional methods according to claim 6.

[Claim 8]The document retrieval method according to claim 7 searching only a character string containing said set-up keyword with said retrieval process out of a document of said retrieval object when an unknown word is not detected in said unknown word detection process.

[Claim 9]A storage in which computer reading is possiblewherein a program code which operates a computer as the document retrieval system according to any one of claims 1 to 5 is stored.

[Claim 10]A storage in which computer reading is possiblewherein a program code realizable by computer is stored in the keyword additional methods according to

claim 6.

[Claim 11]A storage in which computer reading is possiblewherein a program code realizable by computer is stored in the document retrieval method according to claim 7 or 8.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention]This invention relates to the document retrieval system which outputs efficiently the document which contains the keyword or is related based on the keyword inputted as a retrieval object.

[0002]

[Description of the Prior Art]Conventionallythe art of searching the document containing the keyword inputted by the operator out of two or more documents memorized beforehand is proposed using the computer.

[0003]The dictionary in which other words associated to a certain keyword are beforehand registered in recent years. Prepare (the following and a key word dictionary)acquire the word which is relevant to the keyword inputted by the operator from the key word dictionary concernedand search of a document is facedIn addition to the inputted keywordthe art of combining the word acquired from the key word dictionary concerned as other keywordsand using it is also proposed.

[0004]

[Problem(s) to be Solved by the Invention]When the keyword inputted into the key word dictionary by the operator is contained according to the document-retrieval processing using the above-mentioned key word dictionaryeven if the keyword concerned is not contained in the document of a retrieval objectThe document which is actually relevant can also be hit and a document retrieval can be performed efficiently.

[0005]Howeverwhen the inputted keyword is a strange word (the followingunknown word) which is not registered into a key word dictionaryonly the document which contains the keyword directly can be hit. For this reasonwhen a neologism and an unknown word are inputted as a keywordan efficient document retrieval cannot be performed.

[0006]Thenthis invention aims at offer of the storage in which the keyword additional methods to the document retrieval system which performs an efficient document retrieval also to a strange keywordand its devicea document retrieval methodand computer reading are possible.

[0007]

[Means for Solving the Problem]In order to attain the above-mentioned purposea document retrieval system concerning this invention is characterized by the following composition.

[0008]That is this invention is characterized by that a document retrieval system with which a character string containing the keyword is searched out of a document of a retrieval object comprises the following based on a set-up keyword. An unknown word detection means to judge whether it is an unknown word by which a word which extracted a word and was extracted is not registered into the 1st dictionary (key word dictionary) from an input sentence document.

A word judged to be an unknown word by said unknown word detection means. A correlation means to draw up the 2nd dictionary in which a word located the front or behind this word was associated. [in said input sentence document] By referring to said 2nd dictionary (correlation dictionary) based on a set-up keywordA keyword adding means which sets additionally a word related with this keyword as a new keywordA search means to search a character string containing at least one of these keywords out of a document of said retrieval object according to two or more keywords developed by said keyword adding means.

[0009]In a suitable embodiment said unknown word detection means and said correlation means are performed to a document of said retrieval object at a computer's ticket or stage amphiboles separately from execution timing of said keyword adding means and said search means.

[0010]In order to attain the above-mentioned purpose keyword additional methods to a document retrieval system concerning this invention are characterized by the following composition.

[0011]Namely it is a document retrieval method which searches a character string containing the keyword out of a document of a retrieval object based on a set-up keywordAn unknown word detection process which extracts a word and judges whether it is an unknown word by which an extracted word is not registered into the 1st dictionary (key word dictionary) from an input sentence documentA correlation process of drawing up the 2nd dictionary (correlation dictionary) in which a word judged to be an unknown word in said unknown word detection process and a word located the front or behind this word were associated[in said input sentence document] By referring to said 2nd dictionary based on a set-up keywordit has like keyword additional processing which sets additionally a word related with this keyword as a new keyword.

[0012]According to two or more keywords which are document retrieval methods and were developed with the above-mentioned keyword additional methods a character string containing at least one of these keywords is searched out of a document of said retrieval object.

[0013]It is characterized by a storage with which a program code which realizes keyword additional methods to the above-mentioned document retrieval system and its device and a document retrieval method by computer is stored and in which computer reading is possible.

[0014]

[Embodiment of the Invention]Hereafter one embodiment of this invention is described with reference to drawings. The composition of the computer which first

can realize the document retrieval system concerning this invention is explained with reference to drawing 13.

[0015]Drawing 13 is a block lineblock diagram of the computer which can apply this invention.

[0016]22 are a keyboard which is an input means of the display of CRT etc.the keyword for search in 23a documentetc.etc. among a figure. 24 is ROM which has memorized the boot program etc. 25 is RAM which stores a various processing result temporarily. 26 is memory storage such as a hard disk drive (HDD) which memorizes the database which consists of two or more documents (document data) used as a retrieval objectthe program which realizes document-retrieval processing mentioned lateretc. 27 is a communication interface which can receive document data etc. from an external device. And 28 is a printer which prints a processing result etc. Each of these composition is connected via the internal bus 29and CPU(central processing unit) 21 controls the whole computer as a document retrieval system according to the program memorized to the memory storage 26.

[0017]The database (D/B) which consists of two or more documents (document data) is memorized by the external deviceand may be composition which accesses the external device via the communication interface 27.

[0018]Nextthe procedure of the document retrieval realized by computer shown in drawing 13 is explained. The document-retrieval processing concerning this embodiment is divided roughlyand consists of word-index creation processing (drawing 1) in which it explains belowand document-retrieval processing including a keyword addition (deployment) process.

[0019]Hereword-index creation processing is a commuter's ticket or positioning as a maintenance performed irregularly for the document group in D/B memorized by the memory storage 26 according to directions of an operator that the correlation dictionary mentioned later should be updated. On the other handwhen a keyword is inputted by an operatordocument-retrieval processingReferring to the correlation dictionary already drawn up by the word-index creation processing concerned (updating)and/or the inputted keyword is includedit is the processing which actually extracts a relevant document (character string) out of the document group in D/B memorized by the memory storage 26.

[0020]In [that explanation should be made easy in this embodiment] the above-mentioned word-index creation processing (drawing 1)Mention "TORUSHIO" as an example as a word which is not registered into a correlation dictionaryand the keyword inputted is written with "TORUSHIO" also in the document-retrieval processing (drawing 2) explained belowAlthough the attention part in the documents 1 thru/or 3 of drawing 12 which illustrates the search results of document-retrieval processing is the same as the attention part (portion the character string was actually indicated to be) in the documents 1 thru/or 3 of drawing 3 observed in word-index creation processingIt is from the convenience of explanation and a desirable document is actually chosen from the inside of D/B of a retrieval object according to the keyword inputted.

[0021]<Word-index creation processing> drawing 1 is a flow chart which shows the word-index creation processing in this embodiment.

[0022]In the figurethe word which constitutes the character string is extracted from the character string of these documents by word extraction processing of Step S1 by making applicable to extraction a document group (documents 1 thru/or 3) which is illustrated to drawing 3. When extracting a word in this stepthe word already registered into the key word dictionary searches whether it is contained in each character string for word extraction by referring to a key word dictionary which is illustrated to drawing 6 memorized beforehand.

[0023]When the word which is not registered into the key word dictionary concerned is contained in the character string for word extraction at this timethe character type etc. which continue the front or behind that word are judgedand the word concerned is extracted. If the document 1 shown in drawing 3 is more specifically made into an exampleby referring to a key word dictionaryThe word "TORUSHIO" which wordssuch as unillustrated a "supervisor" a "teaching method" and a "brains target"are extracted by drawing 6and is not registered into the key word dictionary concerned is judged as a word which katakana followsand is extracted out of the character string of the document 1.

[0024]In the unknown word detection processing of Step S2in each word extracted at Step S1It judges whether the unknown word which is not registered into a key word dictionary existsprogresses to Step S3 by this judgment at the time (the word "TORUSHIO" is an unknown word in the case of the above-mentioned example) of YES (those with an unknown word)and progresses to Step S5 at the time of NO (with no unknown word).

[0025]In correlation creation processing of Step S3the unknown word detected at Step S2 and the word which exists the front or behind the unknown word concerned are associated. [in the character string from which the unknown word was extracted] In the case of the above-mentioned examplethe "supervisor" which follows the word is related with "TORUSHIO" detected as an unknown word in the document 1 shown in drawing 3. When it progresses to step S4 when correlation is able to be created at this stepand it cannot createit progresses to Step S5.

[0026]In the correlation storage processing of step S4the correlation created at Step S3 is saved as a correlation dictionary at RAM25 or the memory storage 26. By performing such correlation for each character string for word extractionin the case of three documents illustrated to drawing 3as a correlation dictionary is shown in drawing 9after a "supervisor" and "Japan" have been related with an unknown word "TORUSHIO"it is newly saved.

[0027]In word-index creation processing of Step S5a word index is created by summarizing each word (an unknown word is included) extracted at Step S1 for every document. In the case of the above-mentioned examplethree word indices shown in drawing 5 are created by each word extracted from the documents 1 thru/or 3 shown in drawing 3. These word indices are saved in the state where it was related within the document data which became the origin of creation of the

word indexand D/Band are referred to in Step S15 of the document-retrieval processing (drawing 2) mentioned later.

[0028]<Document-retrieval processing> drawing 2 is a flow chart which shows the document-retrieval processing in this embodiment.

[0029]In the figurethe keyword inputted from keyboard 23 grade by an operator is acquired by the retrieval key word acquisition processing of Step S11. Hereas shown in drawing 4TORUSHIOshall be acquired as a keyword.

[0030]In correlation word retrieval processing of Step S12. It is searched whether the keyword inputted at Step S11 is already registered into the correlation dictionary at presentAs a result of this searchwhen already registeredit progresses to Step S13and when unregisteredit progresses to Step S15 that the document which contains directly the keyword acquired at Step S11 in the document of a retrieval object should be searched. In the case of the above-mentioned examplethe keyword "TORUSHIO" is already registered into the correlation dictionary shown in drawing 9 by the word-index creation processing mentioned above with reference to drawing 1.

[0031]In the acquisition processing of the word corresponding to correlation of Step S13the word (henceforthcorrespondence word) related with the word searched with Step S12 is acquired. Since the keyword "TORUSHIO" can be hit out of a correlation dictionary in Step S12 in the above-mentioned casethe "supervisor" and two words of "Japan" which are related with "TORUSHIO" are acquired from the correlation dictionary concerned as a correspondence word.

[0032]In the retrieval key word adding processing of Step S14it adds as other keywords which have relevance in the keyword into which the correspondence word acquired at Step S13 was inputted at Step S11. In the above-mentioned casein addition to the keyword "TORUSHIO" of drawing 4 set up by the operator at Step S11as shown in drawing 10two words are newly [a "supervisor" and "Japan"] added as a keyword for search. That isin this embodimentit means that it had been developed by two or more keywords which are relevant to the keyword in addition to "TORUSHIO" which was the original retrieval key word.

[0033]Step S15: Search with this step the document group related with the word index concerned within D/B by referring to the word index memorized at Step S5 of drawing 1 according to at least one keyword prepared by the above-mentioned step.

[0034]That iswhen the keyword inputted at Step S11 relates and it does not register with a dictionarygeneral retrieval processing which searches the document (word index) which contains the keyword directly is performed. In this caseif the inputted keyword is "TORUSHIO"the documents 1 and 2 shown in drawing 12 will be obtained as search results.

[0035]On the other handwhen other keywords are added by each step to Step S14 in addition to the keyword inputted at Step S11If those keywords are "TORUSHIO" a "supervisor" and "Japan"the documents 1 thru/or 3 shown in drawing 12 will be obtained as search results by the word index by which those keywords are contained.

[0036] Thus by developing the retrieval key word inputted by the operator using the correlation dictionary drawn up by word-index processing of drawing 1 according to this embodiment Efficient document-retrieval processing can be performed although the keyword inputted by the operator is not included directly the document with which it is relevant to operation can also be hit and a document retrieval can be performed efficiently.

[0037] By referring to the key word dictionary (for example drawing 6) memorized beforehand and an unillustrated dictionary before searching D/B with Step S15 By performing keyword development processing known conventionally drawing up a keyword deployment dictionary (refer to the example on the right-hand side of drawing 7) and referring to the keyword deployment dictionary The keyword (drawing 11) developed at Step S14 mentioned above can be developed in many keywords as shown in drawing 11. Therefore if a document retrieval is performed using the keyword of these plurality although the inputted keyword is not included directly the document which is actually relevant can be hit by higher establishment and operativity improves.

[0038]

[Other embodiments] Even if it applies this invention to the system which comprises two or more apparatus (for example a host computer an interface device a reader a printer etc.) it may be applied to the devices (for example a copying machine a facsimile machine etc.) which consist of one apparatus.

[0039] The purpose of this invention the storage (or recording medium) which recorded the program code of the software which realizes the function of an embodiment mentioned above It is attained also when a system or a device is supplied and the computer (or CPU and MPU) of the system or a device reads and executes the program code stored in the storage. In this case the function of an embodiment which the program code itself read from the storage mentioned above will be realized and the storage which memorized that program code will constitute this invention. By executing the program code which the computer read A part or all of processing that the operating system (OS) etc. which the function of an embodiment mentioned above is not only realized but are working on a computer based on directions of the program code are actual is performed and it is contained also when the function of an embodiment mentioned above by the processing is realized.

[0040] After the program code read from the storage was written in the memory with which the function expansion unit connected to the expansion card inserted in the computer or the computer is equipped Based on directions of the program code a part or all of processing that CPU etc. with which the expansion card and function expansion unit are equipped are actual is performed and it is contained also when the function of an embodiment mentioned above by the processing is realized.

[0041]

[Effect of the Invention] According to this invention explained above offer of the storage in which the keyword additional methods to the document retrieval system

which performs an efficient document retrieval also to a strange keyword and its device a document retrieval method and computer reading are possible is realized.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1]It is a flow chart which shows the word-index creation processing in this embodiment.

[Drawing 2]It is a flow chart which shows the document-retrieval processing in this embodiment.

[Drawing 3]It is a figure which illustrates the document group referred to for creation of a correlation dictionary.

[Drawing 4]It is a figure which illustrates a correlation dictionary.

[Drawing 5]It is a figure showing the example of creation of a word index.

[Drawing 6]It is a figure showing the example of a key word dictionary.

[Drawing 7]It is a figure which illustrates a keyword deployment dictionary.

[Drawing 8]It is a figure which illustrates search results.

[Drawing 9]It is a figure showing the example of creation of a correlation dictionary.

[Drawing 10]It is a figure showing the example of deployment of a keyword.

[Drawing 11]It is a figure which illustrates the case where the example of deployment of the keyword shown in drawing 10 is further developed with reference to a keyword deployment dictionary (drawing 7).

[Drawing 12]It is a figure which illustrates the document group which is search results.

[Drawing 13]It is a block lineblock diagram of the computer which can realize this invention.

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2002-108887

(P2002-108887A)

(43)公開日 平成14年4月12日(2002.4.12)

(51)Int.Cl.⁷

G 0 6 F 17/30

識別記号

2 1 0

1 7 0

3 2 0

F I

G 0 6 F 17/30

テマコート⁷(参考)

2 1 0 A 5 B 0 7 5

1 7 0 A

3 2 0 C

審査請求 未請求 請求項の数11 OL (全8頁)

(21)出願番号

特願2000-299971(P2000-299971)

(22)出願日

平成12年9月29日(2000.9.29)

(71)出願人 000001007

キヤノン株式会社

東京都大田区下丸子3丁目30番2号

(72)発明者 中越 治樹

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

(74)代理人 100076428

弁理士 大塚 康徳 (外2名)

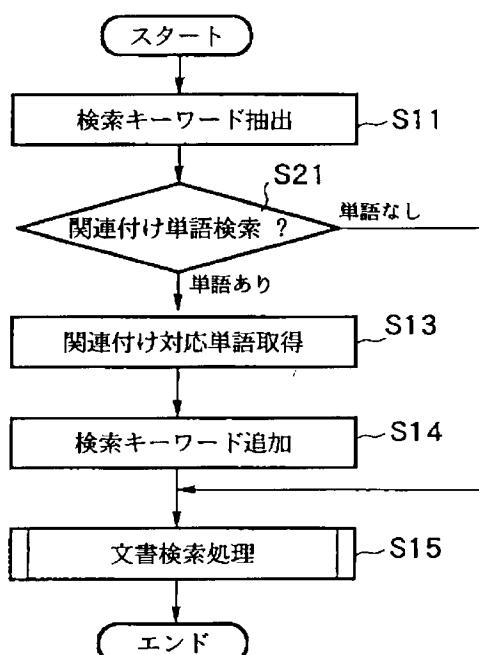
Fターム(参考) 5B075 ND03 NK02 NK31 PP22

(54)【発明の名称】 文書検索装置、その装置へのキーワード追加方法、文書検索方法及びコンピュータ読み取り可能な記憶媒体

(57)【要約】

【課題】 未知のキーワードに対しても効率的な文書検索を行う。

【解決手段】 設定したキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索するに際して、設定されたキーワードに基づいて、関連付け辞書を参照することにより、当該キーワードに関連付けされている単語を他のキーワードとして追加設定し、それら複数のキーワードに従って、検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索する(S11-S15)。ここで、関連付け辞書は、入力文書から抽出した単語が、予め記憶されているキーワード辞書に登録されていない未知語である場合に、その入力文書における当該未知語の前方または後方に位置する単語とを関連付けした辞書であり、S11-S15に先立って作成される。



【特許請求の範囲】

【請求項1】 設定されたキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索装置であって、
入力文書から単語を抽出し、抽出した単語が第1の辞書に登録されていない未知語か否かを判断する未知語検出手段と、
前記未知語検出手段によって未知語であると判断された単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書を作成する関連付け手段と、
設定されたキーワードに基づいて前記第2の辞書を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加手段と、を備えることを特徴とする文書検索装置。

【請求項2】 更に、前記キーワード追加手段によって展開された複数のキーワードに従って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索する検索手段を備えることを特徴とする請求項1記載の文書検索装置。

【請求項3】 前記検索手段は、前記未知語検出手段によって未知語が検出されないときに、前記検索対象の文書の中から、前記設定されたキーワードを含む文字列だけを検索することを特徴とする請求項2記載の文書検索装置。

【請求項4】 前記未知語検出手段及び前記関連付け手段は、前記キーワード追加手段及び前記検索手段の実行タイミングとは別個に、前記検索対象の文書に対して定期または不定期に実行されることを特徴とする請求項1または請求項2記載の文書検索装置。

【請求項5】 前記キーワード追加手段は、前記設定されたキーワード、前記第2の辞書を参照することによって追加されたキーワード、並びに前記第1の辞書に基づいて、それらキーワードを更に追加設定する手段を含むことを特徴とする請求項1記載の文書検索装置。

【請求項6】 設定したキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索装置へのキーワード追加方法であって、
入力文書から単語を抽出し、抽出した単語が第1の辞書に登録されていない未知語か否かを判断する未知語検出手程と、

前記未知語検出手程にて未知語であると判断した単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書を作成する関連付け工程と、

設定されたキーワードに基づいて前記第2の辞書を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加工程と、を有することを特徴とするキーワード追加方法。

【請求項7】 請求項6記載のキーワード追加方法によって展開した複数のキーワードに従って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索することを特徴とする文書検索方法。

【請求項8】 前記検索工程では、前記未知語検出手程にて未知語を検出しなかったときに、前記検索対象の文書の中から、前記設定されたキーワードを含む文字列だけを検索することを特徴とする請求項7記載の文書検索方法。

【請求項9】 請求項1乃至請求項5の何れかに記載の文書検索装置としてコンピュータを動作させるプログラムコードが格納されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項10】 請求項6記載のキーワード追加方法をコンピュータによって実現可能なプログラムコードが格納されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項11】 請求項7または請求項8記載の文書検索方法をコンピュータによって実現可能なプログラムコードが格納されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】**【0001】**

【発明の属する技術分野】 本発明は、検索対象として入力されたキーワードに基づいて、そのキーワードを含むまたは関係する文書を効率良く出力する文書検索装置に関する。

【0002】

【従来の技術】 従来より、コンピュータを用いて、オペレータによって入力されたキーワードを含む文書を、予め記憶されている複数の文書の中から検索する技術が提案されている。

【0003】 また、近年においては、あるキーワードに対して関連付けされた他の単語が予め登録されている辞書（以下、キーワード辞書）を用意しておき、オペレータによって入力されたキーワードに関連性のある単語を当該キーワード辞書から取得し、文書の検索に際しては、入力されたキーワードに加え、当該キーワード辞書から取得した単語を他のキーワードとして併せて利用する技術も提案されている。

【0004】

【発明が解決しようとする課題】 上記のキーワード辞書を利用する文書検索処理によれば、そのキーワード辞書にオペレータによって入力されたキーワードが含まれる場合、検索対象の文書に当該キーワードが含まれていなくても、実際には関連性がある文書をもヒットすることができ、文書検索を効率的に行うことができる。

【0005】 しかしながら、入力されたキーワードがキーワード辞書に登録されていない未知の単語（以下、未

知語)である場合には、そのキーワードを直接的に含む文書しかヒットすることはできない。このため、キーワードとして新語や未知語が入力された場合には、効率的な文書検索を行うことができない。

【0006】そこで本発明は、未知のキーワードに対しても効率的な文書検索を行う文書検索装置及びその装置へのキーワード追加方法、文書検索方法及びコンピュータ読み取り可能な記憶媒体の提供を目的とする。

【0007】

【課題を解決するための手段】上記の目的を達成するため、本発明に係る文書検索装置は、以下の構成を特徴とする。

【0008】即ち、設定されたキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索装置であって、入力文書から単語を抽出し、抽出した単語が第1の辞書(キーワード辞書)に登録されていない未知語か否かを判断する未知語検出手段と、前記未知語検出手段によって未知語であると判断された単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書を作成する関連付け手段と、設定されたキーワードに基づいて前記第2の辞書(関連付け辞書)を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加手段と、前記キーワード追加手段によって展開された複数のキーワードに従って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索する検索手段とを備えることを特徴とする。

【0009】好適な実施形態において、前記未知語検出手段及び前記関連付け手段は、前記キーワード追加手段及び前記検索手段の実行タイミングとは別個に、前記検索対象の文書に対して定期または不定期に実行される。

【0010】また、上記の同目的を達成するため、本発明に係る文書検索装置へのキーワード追加方法は、以下の構成を特徴とする。

【0011】即ち、設定したキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索方法であって、入力文書から単語を抽出し、抽出した単語が第1の辞書(キーワード辞書)に登録されていない未知語か否かを判断する未知語検出工程と、前記未知語検出工程にて未知語であると判断した単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書(関連付け辞書)を作成する関連付け工程と、設定されたキーワードに基づいて前記第2の辞書を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加工程を有することを特徴とする。

【0012】また、文書検索方法であって、上記のキーワード追加方法によって展開した複数のキーワードに従

って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索することを特徴とする。

【0013】更に、上記の文書検索装置、その装置へのキーワード追加方法、並びに文書検索方法を、コンピュータによって実現するプログラムコードが格納されている、コンピュータ読み取り可能な記憶媒体を特徴とする。

【0014】

【発明の実施の形態】以下、本発明の一実施形態を図面を参照して説明する。はじめに、本発明に係る文書検索装置を実現可能なコンピュータの構成を図13を参照して説明する。

【0015】図13は、本発明を適用可能なコンピュータのブロック構成図である。

【0016】図中、22は、CRT等のディスプレイ、23は検索用のキーワードや文書等の入力手段であるキーボードである。24は、ブートプログラム等を記憶しているROMである。25は、各種処理結果を一時記憶するRAMである。26は、検索対象となる複数の文書(文書データ)からなるデータベースや、後述する文書検索処理を実現するプログラム等を記憶するハードディスクドライブ(HDD)等の記憶装置である。27は、外部装置から文書データ等を受信可能な通信インターフェースである。そして28は、処理結果等を印刷するプリンタである。これらの各構成は、内部バス29を介して接続されており、CPU(中央演算処理装置)21は記憶装置26に記憶したプログラムに従って、文書検索装置としてのコンピュータの全体を制御する。

【0017】尚、複数の文書(文書データ)からなるデータベース(D/B)は、外部装置に記憶されており、その外部装置に通信インターフェース27を介してアクセスする構成であっても良い。

【0018】次に、図13に示すコンピュータによって実現される文書検索の処理手順について説明する。本実施形態に係る文書検索処理は、大別して、以下に説明する単語インデックス作成処理(図1)と、キーワード追加(展開)工程を含む文書検索処理とからなる。

【0019】ここで、単語インデックス作成処理は、後述する関連付け辞書を更新すべく、例えばオペレータの指示に応じて、記憶装置26に記憶されたD/B内の文書群を対象として、定期または不定期に実行されるメンテナンスとしての位置付けである。一方、文書検索処理は、オペレータによってキーワードが入力されたときに、当該単語インデックス作成処理によって既に作成(更新)されている関連付け辞書を参照しながら、入力されたキーワードを含む及び/または関連性のある文書(文字列)を記憶装置26に記憶されたD/B内の文書群の中から実際に抽出する処理である。

【0020】尚、本実施形態では、説明を容易にすべ

く、上記の単語インデックス作成処理（図1）において、関連付け辞書に登録されていない単語として「トルシオ」を例に挙げ、以下に説明する文書検索処理（図2）においても、入力されるキーワードを「トルシオ」としたため、単語インデックス作成処理において注目した図3の文書1乃至3における注目箇所（実際に文字列が示された部分）と、文書検索処理の検索結果を例示する図12の文書1乃至3における注目箇所とが同じであるが、説明の都合からであり、実際には、入力されるキーワードに応じて、検索対象のD/B内から好みの文書が選択される。

【0021】<単語インデックス作成処理>図1は、本実施形態における単語インデックス作成処理を示すフローチャートである。

【0022】同図において、ステップS1の単語抽出処理では、図3に例示するような文書群（文書1乃至3）を抽出対象として、それら文書の文字列から、その文字列を構成する単語を抽出する。本ステップにおいて単語を抽出する際には、予め記憶されている図6に例示するようなキーワード辞書を参照することにより、そのキーワード辞書に既に登録されている単語が、単語抽出対象の各文字列に含まれているか否かを検索する。

【0023】このとき、単語抽出対象の文字列の中に、当該キーワード辞書に登録されていない単語が含まれていた場合には、その単語の前方または後方に連続する文字種などを判断して当該単語を抽出する。より具体的には、図3に示す文書1を例にすると、キーワード辞書を参照することにより、図6には不図示の「監督」、「指導方法」、「頭脳的」などの単語が抽出され、当該キーワード辞書に登録されていない単語「トルシオ」は、カタカナが連続する単語として判断されて、文書1の文字列の中から抽出される。

【0024】ステップS2の未知語検出処理では、ステップS1で抽出した各単語の中に、キーワード辞書に登録されていない未知語が存在するかどうかを判断し、この判断でYES（未知語有り）のとき（上記の例の場合は、単語「トルシオ」が未知語）にはステップS3に進み、NO（未知語無し）のときにはステップS5に進む。

【0025】ステップS3の関連付け作成処理では、ステップS2で検出された未知語と、その未知語が抽出された文字列における当該未知語の前方または後方に存在する単語とを関連付けする。上記の例の場合、図3に示す文書1において、未知語として検出された「トルシオ」に、その単語に後続する「監督」が関連付けされる。本ステップにて関連付けが作成できたときにはステップS4に進み、作成できないときにはステップS5に進む。

【0026】ステップS4の関連付け保存処理では、ステップS3で作成した関連付けを、関連付け辞書とし

て、RAM25や記憶装置26に保存する。このような関連付けが単語抽出対象の各文字列を対象にして行われることにより、図3に例示する3つの文書の場合には、関連付け辞書に、図9に示すように未知語「トルシオ」に「監督」と「ジャパン」とが関連付けされた状態で新たに保存される。

【0027】ステップS5の単語インデックス作成処理では、ステップS1で抽出された各単語（未知語を含む）を文書毎にまとめることにより、単語インデックスが作成される。上記の例の場合には、図3に示す文書1乃至3から抽出された各単語により、図5に示す3つの単語インデックスが作成される。これらの単語インデックスは、その単語インデックスの作成の元となった文書データとD/B内で関連付けされた状態で保存され、後述する文書検索処理（図2）のステップS15において参照される。

【0028】<文書検索処理>図2は、本実施形態における文書検索処理を示すフローチャートである。

【0029】同図において、ステップS11の検索キーワード取得処理では、オペレータによってキーボード23等から入力されるキーワードを取得する。ここでは、図4に示すように、キーワードとして「トルシオ」が取得されるものとする。

【0030】ステップS12の関連付け単語検索処理では、ステップS11で入力されたキーワードが、関連付け辞書に現時点で既に登録されているか否かを検索し、この検索の結果、既に登録されているときにはステップS13に進み、未登録のときには、検索対象の文書の中にステップS11にて取得したキーワードを直接含む文書を検索すべくステップS15に進む。上記の例の場合は、図1を参照して上述した単語インデックス作成処理により、図9に示す関連付け辞書には、キーワード「トルシオ」が既に登録されている。

【0031】ステップS13の関連付け対応単語の取得処理では、ステップS12で検索された単語に関連付けされている単語（以下、対応単語）を取得する。上記の場合は、ステップS12において、キーワード「トルシオ」を、関連付け辞書の中からヒットすることができるので、当該関連付け辞書から「トルシオ」に関連付けされている「監督」及び「ジャパン」の2つの単語が対応単語として取得される。

【0032】ステップS14の検索キーワード追加処理では、ステップS13で取得した対応単語を、ステップS11にて入力されたキーワードに関連性を有する他のキーワードとして追加する。上記の場合は、ステップS11にてオペレータによって設定された図4のキーワード「トルシオ」に加えて、図10に示すように、「監督」及び「ジャパン」の2語が新たに検索用のキーワードとして追加されている。即ち、本実施形態では、元の検索キーワードであった「トルシオ」に加え、そのキ-

ワードに関連性のある複数のキーワードに展開されたことになる。

【0033】ステップS15：本ステップでは、上記のステップまでに用意された少なくとも1つのキーワードに従って、図1のステップS5にて記憶した単語インデックスを参照することにより、D/B内で当該単語インデックスに関連付けされている文書群を検索する。

【0034】即ち、ステップS11にて入力されたキーワードが関連付け辞書に登録されてなかった場合には、そのキーワードを直接含む文書（単語インデックス）を検索する一般的な検索処理を行う。この場合、入力されたキーワードが「トルシオ」であれば、検索結果として、図12に示す文書1及び2が得られる。

【0035】一方、ステップS14までの各ステップにより、ステップS11にて入力されたキーワードに加え、他のキーワードが追加された場合には、それらのキーワードが「トルシオ」、「監督」及び「ジャパン」であれば、それらのキーワードが含まれる単語インデックスにより、検索結果として、図12に示す文書1乃至3が得られる。

【0036】このように、本実施形態によれば、オペレータによって入力された検索キーワードを、図1の単語インデックス処理によって作成された関連付け辞書を利用して展開することにより、効率的な文書検索処理を行うことができ、オペレータによって入力されたキーワードを直接的には含まないものの、実施には関連性がある文書をもヒットすることができ、文書検索を効率的に行うことができる。

【0037】尚、ステップS15では、D/Bを検索するのに先立って、予め記憶されているキーワード辞書（例えば図6）と不図示の辞書とを参照することにより、キーワード展開辞書（図7の右側の例参照）を作成し、そのキーワード展開辞書を参照しながら、従来より知られているキーワード展開処理を行うことにより、上述したステップS14にて展開済みのキーワード（図11）を、図11に示すように、更に多くのキーワードに展開することができる。従って、これら複数のキーワードを用いて文書検索を行えば、入力されたキーワードを直接的には含まないものの、実際には関連性がある文書を、より高い確立でヒットすることができ、操作性が向上する。

【0038】

【他の実施形態】尚、本発明は、複数の機器（例えばホストコンピュータ、インタフェイス機器、リーダ、プリンタなど）から構成されるシステムに適用しても、一つの機器からなる装置（例えば、複写機、ファクシミリ装置など）に適用してもよい。

【0039】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記

録した記憶媒体（または記録媒体）を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはCPUやMPU）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても、達成される。この場合、記憶媒体から読み出されたプログラムコード自身が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているオペレーティングシステム（OS）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれる。

【0040】さらに、記憶媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張カードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張カードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれる。

【0041】

【発明の効果】以上説明した本発明によれば、未知のキーワードに対しても効率的な文書検索を行う文書検索装置及びその装置へのキーワード追加方法、文書検索方法及びコンピュータ読み取り可能な記憶媒体の提供が実現する。

【図面の簡単な説明】

【図1】本実施形態における単語インデックス作成処理を示すフローチャートである。

【図2】本実施形態における文書検索処理を示すフローチャートである。

【図3】関連付け辞書の作成のために参照する文書群を示す図である。

【図4】関連付け辞書を例示する図である。

【図5】単語インデックスの作成例を示す図である。

【図6】キーワード辞書の例を示す図である。

【図7】キーワード展開辞書を例示する図である。

【図8】検索結果を例示する図である。

【図9】関連付け辞書の作成例を示す図である。

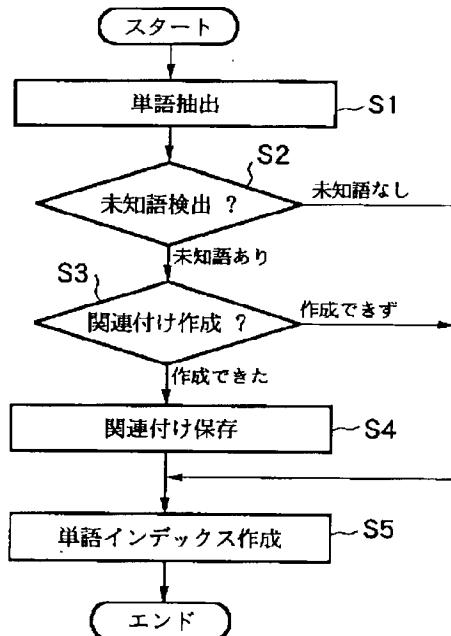
【図10】キーワードの展開例を示す図である。

【図11】図10に示すキーワードの展開例を、キーワード展開辞書（図7）を参照して更に展開した場合を示す図である。

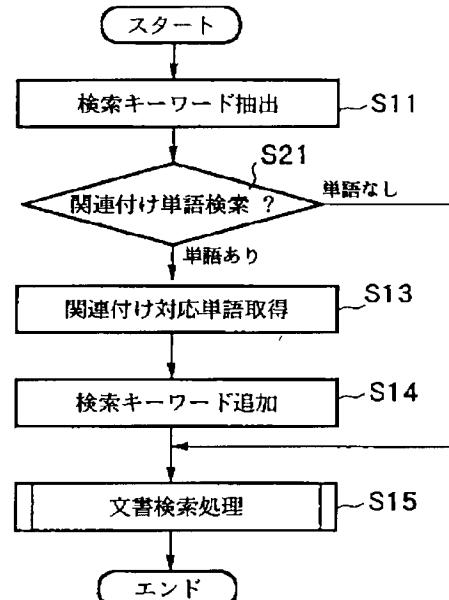
【図12】検索結果である文書群を例示する図である。

【図13】本発明を実現可能なコンピュータのブロック構成図である。

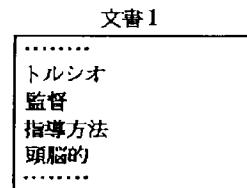
【図1】



【図2】



【図5】



【図4】

トルシオ
<u>検索キーワード</u>

【図11】

トルシオ
監督
ジャパン
日本
Japan
.....

検索キーワード展開

【図6】

.....
トルエン
トルク
トルコ
トルstoi
トルネード
.....

キーワード辞書

【図10】

トルシオ
監督
ジャパン

検索キーワード展開

文書2

.....
トルシオ
ジャパン
初めて
試合
.....

文書3

.....
サッカー
日本
代表
監督
契約
条件
ワールドカップ
出場
.....

単語インデックス群

【図3】

文書1
.....
トルシオ監督の指導方法は、頭脳的であり、
.....

文書2

.....
トルシオジャパンとして、初めての試合が行
われたが、
.....

文書3

.....
サッカー日本代表監督の契約条件としては、
ワールドカップ出場であるが、
.....

【図8】

文書1
.....
トルシオ監督の指導方法は、頭脳的であり、
.....

文書2
.....
トルシオジャパンとして、初めての試合が行
われたが、
.....

検索結果文書群

【図7】

.....	→
ジャパン		日本
.....	→	Japan
トルエン	→	可燃性液体
トルコ	→	トルコ石 トルコ歓談
.....	

キーワード展開辞書

【図12】

文書1
.....
トルシオ監督の指導方法は、頭脳的であり、
.....

文書2

.....
トルシオジャパンとして、初めての試合が行
われたが、
.....

文書3

.....
サッカー日本代表監督の契約条件としては、
ワールドカップ出場であるが、
.....

検索結果文書群

【図9】

.....	→
トルシオ		監督
トンガ	→	ジャパン
.....		王国
.....	

関連付け辞書

【図13】

